Annals of Clinical and Analytical Medicine
Original Research

Use of large language models in radiological reports: A study on simplifying turkish MRI findings

Use of large language models on simplifying turkish MRI reports

Turay Cesur[1], Eren Çamur[2], Yasin Celal Güneş[3]
[1] Department of Radiology, Ankara Mamak State Hospital, Ankara
[2] Department of Radiology, Ministry of Health Ankara 29 Mayis State Hospital, Ankara
[3] Department of Radiology, Yuksek Ihtisas Hospital, Kırıkkale, Türkiye

Abstract

Aim: Advanced Large Language Models (LLMs), like ChatGPT, are known for their human-like expression and reasoning abilities. They are used in many fields, including radiology. This study is pioneering in evaluating and comparing the effectiveness of LLMs in simplifying Magnetic Resonance Imaging (MRI) findings in Turkish.

Material and Methods: In our study, we simplified 50 fictional MRI findings in Turkish language using different LLMs, including ChatGPT-4, Gemini Pro 1.5, Claude 3 Opus and Perplexity. We compared the responses based on Ateşman's readability index and word count. Additionally, three radiologists assessed the medical accuracy, consistency of suggestions, and comprehensibility of the answers, scoring each model on a scale of 1 to 5.

Results: There was no statistically significant difference between the scores of Gemini 1.5 Pro (average: 4.9; median: 5.0), Opus (average: 4.8; median: 5.0), and ChatGPT-4 (average: 4.8; median: 5.0) ($p>0.05$). However, there was a significant difference between the scores of Gemini 1.5 Pro and Perplexity (average: 3.7; median: 4.0) ($p<0.001$).

According to the readability index, Gemini 1.5 Pro had the highest average score of 59.3, which was significantly higher than the other LLMs ($p<0.005$). In terms of word count, ChatGPT-4 used the most words (151.5), while Perplexity used the fewest (88.4).

Discussion: This study is the first to evaluate the ability of LLMs to simplify MRI findings in Turkish. The results suggest that radiologists find these models effective in making radiology reports more understandable. However, additional research is necessary to confirm these findings.

Keywords

Large Language Model, Radiology Reports, Readability, Health Communication

DOI: 10.4328/ACAM.22266 Received: 2024-05-17 Accepted: 2024-07-02 Published Online: 2024-07-10 Printed: 2024-08-01 Ann Clin Anal Med 2024;15(8):586-590
Corresponding Author: Turay Cesur, Department of Radiology, Ankara Mamak State Hospital, Ankara, Türkiye.
E-mail: turaycesur93@gmail.com P: +90 534 875 75 90
Corresponding Author ORCID ID: https://orcid.org/0000-0002-2726-8045
Other Authors ORCID ID: Eren Çamur, https://orcid.org/0000-0002-8774-5800 · Yasin Celal Güneş, https://orcid.org/0000-0001-7631-854X

Introduction

Natural language processing (NLP) tools, especially large language models (LLMs), have gained the ability to generate highly accurate and human-like text through extensive training on large datasets [1]. ChatGPT, a state-of-the-art NLP model released by OpenAI in November 2022, has gained worldwide attention for its human-like expression and reasoning abilities. It has been applied in various fields, including writing, summarization, medical knowledge, and medical education [2, 3].

In the medical field, LLMs have been the focus of many studies and have sparked interest in radiology [4, 5]. Their success in radiological assessments, understanding of radiological guidelines, and contributions to differential diagnosis and decision-making have recently generated significant excitement among radiologists [6, 7].

Radiology reports summarize radiologists' findings and opinions based on imaging studies. They are critical in daily practice. However, medical terminology and key insights in these reports can be hard to understand for patients and physicians from other specialities. LLMs now enable the adaptation, simplification, and translation of professional radiology reports into different languages. This makes the reports comprehensible to individuals without medical knowledge and highlights key points [8, 9, 10]. Successful simplification of radiology reports by LLMs can greatly enhance health communication and clarity for patients and their relatives. To the best of our knowledge, there are no studies on this subject for MRI findings in Turkish yet. Given the lack of studies on this topic in Turkish language, we aimed to evaluate and compare the effectiveness of LLMs, particularly in simplifying Magnetic Resonance Imaging (MRI) findings in Turkish.

Material and Methods

In our study, we evaluated and compared the abilities of large language models (LLMs) such as ChatGPT-4, Gemini 1.5 Pro, Claude 3 Opus, and Perplexity to simplify Turkish MRI findings. Our study only included fictional MRI findings and did not use actual radiology reports, so it did not require ethical board approval. The study design followed the Standards for Reporting of Diagnostic Accuracy Studies (STARD) and the principles outlined in the Declaration of Helsinki [11].

For our study, the authors jointly created 50 fictional Turkish MRI findings used in radiology reports. Care was taken to ensure these findings were common in daily practice and portrayed realistically. The findings included 20 related to neuroradiology, 15 to musculoskeletal radiology, and 15 to abdominal radiology. Table 1 showcases 20 of these findings as examples.

We utilized LLMs named ChatGPT-4, Gemini Pro 1.5, Perplexity, and Claude 3 Opus. We chose these models because of their timeliness, powerful capabilities, and the fact that they come from different companies [4, 5, 12]. The designed findings were entered into each LLM via their respective websites following the prompt, "I will write the findings from the MRI report below. Please explain them in a way that someone without a medical background can understand." in Turkish. Each finding was processed in a new window with default settings used for each model. The study was conducted between April 25 and April 28, 2024.

We analyzed the responses from the LLMs using the Ateşman Readability Index, a well-established Turkish readability measure, to determine readability levels [13]. This index includes the number of syllables of words and the number of words of sentences in its formula [198,825 - (40,175 x number of syllables/number of words) - (2,610 x number of words/number of sentences)]. We used the publicly available and free website "www.okunabilirlikindeksi.com" for this process.

Three authors jointly rated the responses of the LLMs using a Likert scale from 1 to 5, based on medical accuracy, consistency of recommendations, and comprehensibility.

We also measured and compared the word count of the LLMs' responses to assess their word efficiency, which is believed to directly impact readers' reading time. Figure 1 summarizes the workflow of the study.

We used IBM SPSS Version 26 for statistical analyses. We checked data distribution with the Kolmogorov-Smirnov and Shapiro-Wilk tests. The Levene test assessed data variance. Descriptive statistics included minimum, maximum, average, median, standard deviation, interquartile range, and percentages. To find significant relationships between quantitative data in dependent groups, we used the Friedman and Wilcoxon tests. We used Spearman correlation analysis to examine the linearity of correlations between quantitative data.

Results

There was no significant difference between the scores of Gemini 1.5 Pro (average: 4.9; median: 5.0), Opus (mean: 4.8; median: 5.0), and ChatGPT 4 (mean: 4.8; median: 5.0) ($p>0.05$). However, Gemini 1.5 Pro had significantly superior scores compared to Perplexity (mean: 3.7; median: 4.0) ($p<0.001$). Table 3 summarizes the characteristics of each model in our study.

Gemini Pro 1.5 received a score of 5 in 48 out of 50 questions and a score of 4 in 2 questions. Opus scored 5 in 44 questions, 4 in 2 questions, and 3 in 4 questions. ChatGPT 4 scored 5 in

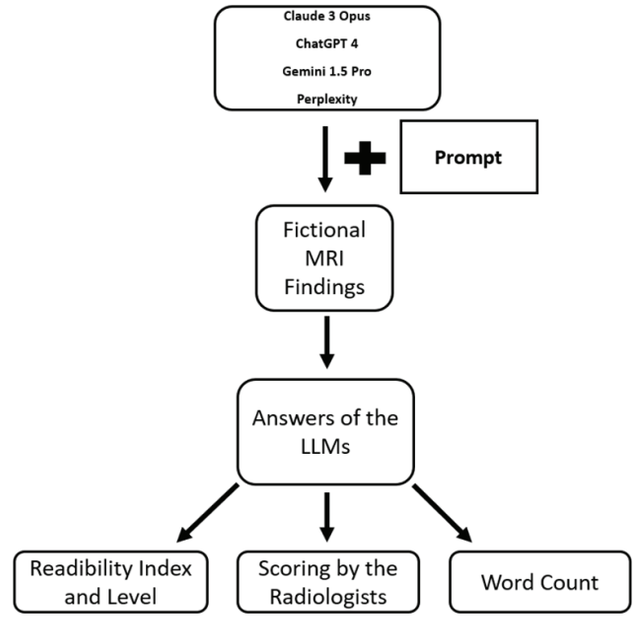

Figure 1. Summarizes the workflow of the study

43 questions and 4 in the remaining 7 questions. The lowest performing LLM in our study was Perplexity, which scored 5 in 9 questions, 4 in 25 questions, 3 in 11 questions, 2 in 4 questions, and 1 in 1 question. Figure 2 shows the box-plot graphs of the scores obtained by the language models.

According to Ateşman's readability index, Gemini 1.5 Pro had the highest average score of 59.3, which was significantly higher than the other LLMs (p<0.005) (Table 3). There was no significant difference between the average scores of ChatGPT 4 (53.85) and Opus (52.86) (p=0.543). Perplexity's average readability index (47.01) was significantly lower than all other LLMs (p<0.001).

ChatGPT 4 had the highest average number of words used (151.5), which was significantly higher than all other language models (p<0.001) (Table 3). Gemini 1.5 Pro, with the second-highest average (121.1), also used significantly more words compared to the other models (p<0.001). Opus (103.1) used significantly more words than Perplexity (88.4) (p=0.008).

There was a linear correlation between the number of words in the designed findings and the number of words produced by Gemini 1.5 Pro (correlation coefficient = 0.706, p<0.000) and ChatGPT 4 (correlation coefficient = 0.585, p<0.000). However, no linear correlation was found for Opus (p=0.224) and Perplexity (p=0.420).

A linear correlation existed between the readability index of the designed findings and the readability indices of the responses from Opus (correlation coefficient = 0.545, p<0.000), Perplexity (correlation coefficient = 0.387, p=0.005), and Gemini 1.5 Pro (correlation coefficient = 0.294, p=0.038). However, no correlation was found between the readability index of the findings and the readability index of ChatGPT 4 (p=0.402).

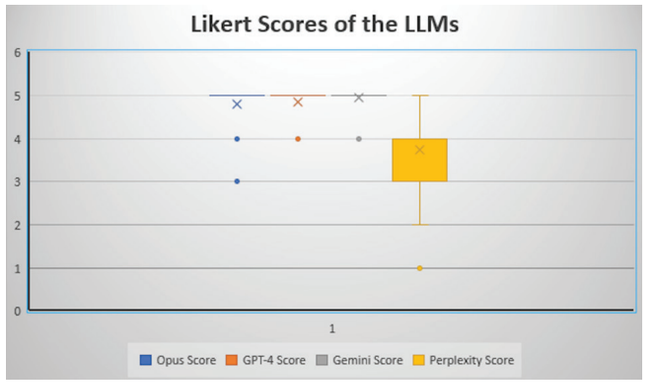

Figure 2. The responses of the LLMs were rated by three authors jointly using a Likert Scale from 1 to 5 based on medical accuracy, consistency of recommendations, and comprehensibility. Box-plot shows the scores of each LLM. x: median, dot: scores showing extreme distribution

Discussion

Our study showed that even fictional, commonly used Turkish findings in MRI reports can be simplified by large language models (LLMs) for people without medical backgrounds. Gemini 1.5 Pro, Claude 3 Opus, and ChatGPT 4 received near-perfect scores from three radiologists for accuracy, consistency, and comprehensibility. Perplexity also scored above average (mean: 3.7/5; median: 4/5).

Gemini 1.5 Pro had the highest Ateşman readability index score (59.3), significantly higher than all other models. ChatGPT 4 (53.85) and Opus (52.86) had mid-range scores. Perplexity had the lowest score (47.01), significantly lower than the other models. We used the Ateşman readability index to measure the readability of the simplified MRI reports generated by the LLMs in Turkish. Ateşman developed this formula in 1997 [13]. It evaluates the readability of Turkish texts based on the average number of syllables per word and words per sentence. Scores range from 1 to 100, with higher scores indicating easier reading

Table 1. The English translations of 20 of the fictional Turkish MRI findings designed for the study are shown

	Translations of the MRI Findings
1	Cavum septum pellucidum variation was observed.
2	The third and both lateral ventricles appear enlarged for the patient's age.
3	A slight asymmetrical enlargement of the right lateral ventricle body compared to the left was noted.
4	A 32 mm retention cyst was observed in the left maxillary sinus.
5	A 14 mm mucosal thickening was observed on the posterior wall of the nasopharynx, significantly narrowing the air column.
6	The third and right lateral ventricles are compressed, with a 9 mm leftward shift of midline structures and minimal uncal herniation on the right observed.
7	An 18 mm encephalomalacic and gliotic area was observed in the right occipital lobe, with T1-weighted hyperintense signals consistent with cortical laminar necrosis.
8	Degenerative spurs are observed at the corners of the vertebral bodies.
9	T2-weighted signal changes and occasional height loss due to degeneration are present in the intervertebral discs at the lumbar level.
10	In the superior-middle section of the posterior labrum, hyperintense signal changes suspicious for degeneration or tear are present.
11	Fluid levels were observed in the locations of both trochanteric bursae, and it has been evaluated as bilateral trochanteric bursitis.
12	The coverage ratio of the femoral head by the left acetabulum has increased (possible pincer-type FAI?).
13	Narrowing of joint spaces, millimetric osteophytic changes, and minimal fluid increases are present (possible osteoarthritis?).
14	Diffuse and significant signal loss consistent with steatosis was observed in the out-of-phase images of the liver.
15	A 10 mm accessory splenic tissue was observed near the inferior border of the spleen.
16	A 13x10 mm solid nodular lesion in the medial crus of the left adrenal gland, showing a signal intensity index of 47% with signal loss in out-of-phase images, was primarily evaluated as an adenoma.
17	At the hepatic flexure level of the colon, a suspicious wall thickening and enhancement, up to 6 mm at its thickest point, was noted in an approximately 3 cm segment.
18	The craniocaudal length of the right lobe of the liver was measured at 165 mm in the midclavicular line.
19	Diffuse signal intensity loss in the medullary area of the bone structures was noted on T1-weighted sequences (bone marrow reconversion? malignancy?).
20	Loss of integrity and increased signal at the talar attachment site of the ATFL was detected (partial tear?).

[13]. Ateşman highlighted that a text's success depends on both readability and comprehensibility. Readability is evaluated with quantitative data, while comprehensibility is assessed qualitatively using the content of the text [13]. We evaluated the readability of the responses with the Ateşman index and their comprehensibility using a Likert scale. We acknowledge that ratings by individuals without medical backgrounds would provide more valuable insights into comprehensibility.

ChatGPT 4 used significantly more words (151.5) than the other models, indicating that its responses would take more time to read. Gemini 1.5 Pro followed with 121.1 words, then Opus with 103.1 words, and finally Perplexity with 88.4 words. We evaluated the word counts of the responses, considering the relationship between the number of words and the time required for users to read them.

In a similar study, Jeblick et al. used the prompt, "Explain this medical report to a child using simple language," with ChatGPT 3.5 for three fictional radiology reports [8]. Fifteen radiologists rated the responses based on accuracy, comprehensiveness, and potential harm to the patient. Almost all responses were rated as accurate, comprehensive, and unlikely to cause harm [8]. Our study was conducted entirely in Turkish and examined various LLMs in comparison to ChatGPT 3.5. This approach allowed us to evaluate the performance of other LLMs as well.

In a similar study, Schmidt et al. used ChatGPT 3.5 to simplify knee MRI findings of varying complexity (simple, moderate, and complex) with five different prompts [9]. Four doctors (two orthopedists and two radiologists) and 20 patients evaluated the simplified reports. The doctors rated the reports as "neutral" for informativeness but agreed they were "good" in accuracy and comprehensibility, posing no harm to patients. Patients felt better after understanding the simplified reports [9].

Lyu et al. examined 62 thorax CT and 76 brain MRI reports [10]. Each report had three simplified versions based on different prompts: making the report easier to understand, providing patient advice, and offering healthcare professional recommendations [10]. Two radiologists rated these simplified reports on the overall score, comprehensiveness, and accuracy. ChatGPT 3.5, using more comprehensible language, scored 4.27/5. ChatGPT 4 produced even better-quality reports. They also explored how different prompts could create varied reports for patients with different education levels and found no significant differences [10]. We did not use such specific prompts. Instead, we compared the baseline responses of the models using the prompt "in a way that someone without a medical background can understand." Studies on how prompt engineering can improve readability levels in Turkish reports are needed to see how they would affect the readability levels of the responses.

Li et al. studied 100 X-ray, ultrasound, CT, and MRI reports [14]. They examined their lengths, Flesch reading ease scores, and Flesch-Kincaid reading levels [14]. They used the prompt, "Explain this radiology report to a patient in layman's terms: <Report Text>." The simplified reports were statistically shorter, easier to read, and at a lower reading level than the originals [14]. We did not use specific commands like making the report shorter, longer, or easier to read. Studies exploring the impact of such commands in Turkish reports through

Table 2. The Ateşman Readability Index and its corresponding readability level are shown

Index	Readability Level
90 - 100	Easily understood by 4th grade and below students
80 - 89	Easily understood by 5th or 6th graders
70 - 79	Easily understood by 7th or 8th graders
60 - 69	Easily understood by 9th or 10th graders
50 - 59	Easily understood by 11th or 12th graders
40 - 49	Easily understood by 13th or 15th year (associate degree) students
30 - 39	Easily understood by bachelor's degree
< 30	Easily understood by postgraduates

Table 3. Descriptive findings of the responses of the large language models are shown

	Gemini 1.5 Pro	Opus	ChatGPT 4	Perplexity
Likert Scores*				
Minimum-Maximum	4.0-5.0	3.0-5.0	4.0-5.0	1.0-5.0
Mean ± SD	4.9 ± 0.1	4.8 ± 0.5	4.8 ± 0.3	3.7 ± 0.9
Median (IQR)	5.0 (0)	5.00 (0)	5.00 (0)	4.0 (1.0)
Ateşman Readability Index				
Minimum-Maximum	35.7 - 77.0	24.5 - 69.7	34.7-71.2	21.1 – 78.2
Mean ± SD	59.31 ± 9.13	52.86 ± 7.87	53.85 ± 10.74	47.01 ± 9.92
Median (IQR)	54.1 (14.2)	47.2 (10.3)	37.8 (7.6)	37.3 (6.2)
Readability Level				
Minimum	7.-8. Grade	9.-10. Grade	7.-8. Grade	7.-8. Grade
Maximum	Bachelor's Degree	Postgraduate	Bachelor's Degree	Postgraduate
Median	11.-12. Grade	11.-12. Grade	11.-12. Grade	13.-15. Grade
Word Count				
Minimum-Maximum	72-283	67-162	57-292	40-187
Mean ± SD	125.1 ± 35.4	103.1 ± 25.5	151.5 ± 51.0	88.4 ± 26.5
Median (IQR)	119.5 (24.5)	98.0 (43.7)	147.0 (34.5	83.5 (20.0)

*Likert Scores: In our study, authors rated the accuracy of the explanations, consistency, and comprehensibility of the suggestions made by the large language models on a scale of 1 to 5.
SD: Standard Deviation, IQR: Interquartile range

prompt engineering would be beneficial for demonstrating how much readability and word count can be improved.

Limitation

Our study is the first to examine the simplification of Turkish MRI findings by LLMs for individuals without medical backgrounds. However, there are limitations. The main limitation is the lack of patient inclusion, so we do not have their opinions on these simplified reports. Future studies should compare patient understanding of standard MRI reports with those simplified by LLMs. This would provide valuable feedback on the comprehensibility and usefulness of the simplified reports from the end-users perspective. Another limitation is that we used only fictional findings related to a single condition, not actual radiology reports. More complex reports covering all relevant findings might yield different results. Lastly, we used only one prompt. Different prompts might produce better or worse results depending on the models' capabilities.

Conclusion

Our study suggests that large language models might effectively simplify Turkish MRI findings, potentially enabling patients to read and understand their MRI reports. This understanding could lead to better patient comprehension of their diagnoses and treatments, possibly resulting in enhanced compliance. Nevertheless, further research is required to address the limitations identified in our study and to validate these preliminary findings.

Scientific Responsibility Statement

The authors declare that they are responsible for the article's scientific content including study design, data collection, analysis and interpretation, writing, some of the main line, or all of the preparation and scientific review of the contents and approval of the final version of the article.

Animal and Human Rights Statement

All procedures performed in this study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or compareable ethical standards.

Funding: *None*

Conflict of Interest

The authors declare that there is no conflict of interest.

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940.

2. Grewal H, Dhillon G, Monga V, Sharma P, Buddhavarapu VS, Sidhu G, et al. Radiology Gets Chatty: The ChatGPT Saga Unfolds. Cureus. 2023;15(6):e40135.

3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.

4. Li D, Gupta K, Bhaduri M, Sathiadoss P, Bhatnagar S, Chong J. Comparing GPT-3.5 and GPT-4 Accuracy and Drift in Radiology Diagnosis Please Cases. Radiology. 2024;310(1):e232411.

5. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's Diagnostic Performance from Patient History and Imaging Findings on the Diagnosis Please Quizzes. Radiology. 2023;308(1):e231040.

6. Yilmaz EC, Belue MJ, Turkbey B, Reinhold C, Choyke PL. A Brief Review of Artificial Intelligence in Genitourinary Oncological Imaging. Can Assoc Radiol J. 2023;74(3):534-547.

7. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: Fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol. 2024;30(2):80-90.

8. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: An exploratory case study on simplified radiology reports. Eur Radiol. 2024;34(5):2817-2825.

9. Schmidt S, Zimmerer A, Cucos T, Feucht M, Navas L. Simplifying radiologic reports with natural language processing: A novel approach using ChatGPT in enhancing patient understanding of MRI results. Arch Orthop Trauma Surg. 2024;144(2):611-618.

10. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art. 2023;6(1):9.

11. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. Radiology. 2015;277(3):826-832.

12. Horiuchi D, Tatekawa H, Shimono T, Walston SL, Takita H, Matsushita S, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology. 2024;66(1):73-79.

13. Ateşman E. Türkçede okunabilirliğin ölçülmesi [Measuring readability in Turkish]. Dil Dergisi. 1997;58(2):71-74.

14. Li H, Moon JT, Iyer D, Balthazar P, Krupinski EA, Bercu ZL, et al. Decoding radiology reports: Potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. Clin Imaging. 2023;101:137-141.

How to cite this article:

Turay Cesur, Eren Çamur, Yasin Celal Güneş. *Use of large language models in radiological reports: A study on simplifying turkish MRI findings. Ann Clin Anal Med 2024;15(8):586-590*